Visualization tools for space-time series analysis with context awareness: Montreal subway case

Florian Toqué, PhD candidate
Wednesday, July 10th 2019
Transidata, Paris

Co-direction: IFSTTAR/Université Paris-Est and Polytechnique Montréal
Supervisor: Étienne Côme¹
Directors: Latifa Oukhellou¹ and Martin Trépanier²

¹IFSTTAR (The French institute of science and technology for transport, development and networks), COSYS-GRETTIA
²Polytechnique Montréal and CIRRELT
Table of contents

1. Montreal case
2. Long-term forecasting
3. Visualization tools
4. Conclusion and future work
Montreal case
Case study (data from Société de transport de Montréal (STM))

1 - Ticketing logs (tap in), smart card OPUS:
   - 3 years: 2015-2017 (aggregation 15 minutes)
   - 68 stations, 4 subway lines (blue, green, orange and yellow)
   - Information about the transport pass (28 types of pass)

2 - Event database:
   - 4 years: 2015-2018
   - Event location encoded by the subway station id
   - Start and approximate end date of the event (available end time in 82% of the events)
   - Manual categorization of the type of event into 10 categories

3 - Incident database:
   - Incident location encoded by the subway station id
   - Start and end date of the incident, cause of the incident
   - Blocking metro doors, person injury, etc.
Long-term forecasting
Problem and motivation

Goal

- Predict the number of passengers entering a public transport network
  - In each station
  - Per day (at every time-step of the day, 15 minutes)
  - One year ahead
  - Take into account the event with the forecasting models

Motivations

- Adapt time table planning with seasonal demand
- Propose specific type of pass during special event

Paper under review in Transportmetrica A journal.
F. Toqué, E. Côme, M. Trépanier, L. Oukhellou. Forecasting of the Montreal Subway Smart Card Entry Logs with Event Data
# Table of contents

1. Montreal case

2. Long-term forecasting
   - Problem and motivation
   - Methodology
   - Experiments
   - Results

3. Visualization tools

4. Conclusion and future work
**Methodology (1/2)**

### Input data

- **D1**: basic temporal features (name of the day (1-7) and the month (1-12))
- **D2**: advanced temporal features (name of the day (1-7), month (1-12), school holiday (0-1), Christmas holiday (0-1), holiday (0-1), 24 and 31 of December (0-1), renovation at station Beaubien (0-1))
- **D3**: advanced temporal features + event features
- **D4**: advanced temporal features + event features + category of the event

### Models and input data

The models are univariate, **1 model per time series** (68 stations = 68 models).

Here we compare models with different inputs:

- **2 Historical Average**: HA 1 and HA 2 with inputs D1 and D2
- **4 Random Forest**: RF 1, RF 2, RF 3 and RF 4 with inputs D1, D2, D3 and D4
Methodology (2/2): Trend Factor

- These forecasting methods do not take into account the global trend of the number of passengers from year-to-year.
- In order to take this trend into account in the forecast we have multiplied the forecasted passenger demand by the following trend factor for each station:

\[
\text{trend_factor}_{2015-2016}(s) = \frac{\frac{1}{T_{2016}} \sum_{t_1=0}^{T_{2016}} x_{2016}^{t_1}(s)}{\frac{1}{T_{2015}} \sum_{t_2=0}^{T_{2015}} x_{2015}^{t_2}(s)}
\]

(1)

where

- \(x\) = Number of passengers
- \(T_y\) = Number of time step of year \(y\), with \(t \in T\)
- \(s\) = Station \(s\)

This trend factor corresponds to the ratio of the average passenger demand per timestep of 2016 divided by 2015 per station.
We used the **Root Mean Square Error**, the **Mean Absolute Percentage Error at v** and the **Mean Absolute Error** to compare the models.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{2}
\]

\[
\forall y_i > v, \ MAPE_{@v} = \frac{100}{n} \times \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{4}
\]
Training and testing sets

Figure 1: Long-term experiment data set.
## Results during all the test period (2017)

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE@150</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA 1 (Day)</td>
<td>50.36</td>
<td>21.47</td>
<td>15.28</td>
</tr>
<tr>
<td>HA 2 (Cal)</td>
<td>44.31</td>
<td>19.16</td>
<td>13.84</td>
</tr>
<tr>
<td>RF 1 (Day)</td>
<td>50.39</td>
<td>21.32</td>
<td>15.17</td>
</tr>
<tr>
<td>RF 2 (Cal)</td>
<td>41.35</td>
<td>18.19</td>
<td>13.20</td>
</tr>
<tr>
<td>RF 3 (Cal + event)</td>
<td>39.66</td>
<td>17.99</td>
<td>13.16</td>
</tr>
<tr>
<td>RF 4 (Cal + event + cat.)</td>
<td><strong>38.53</strong></td>
<td><strong>17.88</strong></td>
<td><strong>13.13</strong></td>
</tr>
</tbody>
</table>

- Best long-term results are obtained by the **RF 4** model with the advanced temporal features of the day, the event and the event category features.
- RF 4 model succeeds in generalizing and do not overfit.
Results during event test set period over 17 stations

<table>
<thead>
<tr>
<th>Period without event (2017) on the 17 stations that host event in 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RF 4 (Cal + event + cat.)</strong></td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sub test set (2017) - Event period on the 17 stations that host event in 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RF 1 (Day)</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>RF 2 (Cal)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>RF 3 (Cal + event)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>RF 4 (Cal + event + cat.)</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

- Analysis of the results during event period over all the stations
- Models perform **less accurate forecasting during the event period** (21.07% of MAPE against 13.12% during the period without event)
- The **best long-term model** is the **RF 4** model
Visualization tools
Problems and motivation

Goals:

- Better understand errors in order to improve the model and improve the database
- Understand the impact of disturbance on the different stations of the network

Problem:

- Large number of time-series to analyze (365 days * 68 stations)

Questions we want to solve:

- When do we make errors?
- Why some periods are harder to predict?
- How long is the impact of disturbance? Which stations are impacted?
Two visualization tools: Time and Space

We have developed two types of visualization to analyze the residues of the forecasting:

Time Visualization

Space Visualization
Table of contents

1. Montreal case

2. Long-term forecasting

3. Visualization tools
   - Problem and motivation
   - Time visualization tool
   - Space visualization
   - Event example
   - Incident example

4. Conclusion and future work
Time visualization goals

We have created a time visualization tool with **two time scale resolutions**:

**Year resolution**

- Error aggregation per couple day/station
- Allows to easily spot when the models do big mistakes (day/station) over one year

**Day resolution**

- Residues per timestep/station couple
- Allows to easily spot when the models do big mistakes (timestep/station)
- Allows to know how long are the disturbances
- Allows to understand the impact of disturbances on the different stations
Time visualization documentation

1. Selection of the model residues, the type of normalization, which type of event to show and if we want to filter the visualization on events only

2. Rows: Station name, Column: Day or Timestep (on click, go to the visualization of the day)

3. Selection of the date, and the type of visualization Year or Day

4. Heatmap of the forecasting residues, green = event color. On click: open the information view of the selected day/station
Table of contents

1. Montreal case

2. Long-term forecasting

3. Visualization tools
   - Problem and motivation
   - Time visualization tool
   - Space visualization
   - Event example
   - Incident example

4. Conclusion and future work
We have created a space visualization tool:

- Shows residues values per couple timestep/station depicted in a map
- Allows to analyze the impact of event over the different station of the network
- Allows to select the date and the timestep to visualize

Javascript libraries:

- Mapbox.gl for the map and the metro lines
- D3.js (open source) for the stations
Space visualization documentation

1. Selection the model residues, the type of normalization
2. Selection of the date
3. Player and time slider to change timestep
4. Station, residues (fill color) and event (stroke color)
5. Select style of the map
Day/station information view

1. Events information (start/end datetime, cause, category)
2. Observation and prediction of the different models
3. Link to space visualization on the same date
Table of contents

1. Montreal case

2. Long-term forecasting

3. Visualization tools
   - Problem and motivation
   - Time visualization tool
   - Space visualization
   - Event example
   - Incident example

4. Conclusion and future work
RF4 model: tries to take into account calendar and event features

1. Information about the event: Concert of Bruno Mars
2. Increase of the passenger demand, model rf4 succeed to predict the high increase of the passenger demand
Day example, model rf4, event

1. Event starting at 8pm, Increase of the passenger demand 3 hours after
2. An other station is impacted by the event
1. Montreal case

2. Long-term forecasting

3. Visualization tools
   - Problem and motivation
   - Time visualization tool
   - Space visualization
   - Event example
   - Incident example

4. Conclusion and future work
Year example, model rf4, incident

Select couple day/station with high value of residues filtered on the presence of incident

1. Residues normalization per station
2. Residues with high value and incident the February 15th 2017
1. Incident: person injury at 3pm (train has to stop)
2. High decrease of the ticketing logs (access to the station closed)
1. We can see the different stations impacted by the event and the duration of the disturbance.
1. Incident: person injury (train has to stop), starts at 3:00pm
1. At 3:15pm, many stations are impacted with less passenger than usual (the station maybe closed, service breakdown has been announced)
1. At 3:30pm, stations are impacted with more passengers than usual (the prediction). Passengers are trying to use another line (orange line).
Conclusion and future work
Conclusion of the forecasting

- Long-term models succeed to take into account the event data and achieve accurate predictions one year ahead
- Event categorization is helpful for the forecasting models
- **Future work**: Improve the trend factor between the different years of study
- Create short-term forecasting models with event/incident data as inputs
Conclusion of the visualization

- It is possible to analyze impact of events
- Easier to spotlight period hard to predict and to analyze impact of event over the network
- The visualization allows to spotlight possible event that are not in the database
- **Futur work**: Complete the database and train again the model to see the difference in the forecasting results
- Add weather information to the forecasting/visualization
Thank you for your attention
Any questions?

Research project co-financed by the NSERC, Thalès, Cortex média, with the collaboration of RTC, STL, STM (transit authorities in the province of Quebec)